A Study on Human-Generated Tag Structures to Inform Tag Cloud Layout

Daniela Oelke
Ubiquitous Knowledge Processing Lab
German Institute for Educational Research
Solmsstraße 73
60486 Frankfurt am Main, Germany
oelke@dipf.de

Iryna Gurevych
Ubiquitous Knowledge Processing Lab
German Institute for Educational Research,
Frankfurt & Technische Universität Darmstadt
Hochschulstraße 10
64289 Darmstadt, Germany
gurevych@ukp.informatik.tu-darmstadt.de

ABSTRACT

Tag clouds are popular features on web pages, not only to support browsing but also to provide an overview over the content of the page or to summarize search retrieval results. Commonly, the arrangement of tags is based on a random layout or an alphabetic ordering of the tags. Previous research suggests to further structure the tag clouds according to semantics, typically employing cooccurrence-based relations to assess the semantic relatedness of two tags. Regarding the layout of the resulting structure, a wide variety of representations has been proposed. However, only few papers motivate their design choice or evaluate its performance from the perspective of a user, leaving it open if the approach answers the users' expectations. In this paper we present the results of a study in which we observed how humans structure user-generated tags of a social bookmarking system given the task that the resulting layout should provide a quick overview over a search retrieval result. We examine the participants' layouts based on the final arrangement of tags and a detailed interview conducted after the task. Thereby, we analyze and characterize the different term relations employed as well as the higher-level structures generated. The deeper understanding of what criteria are considered important by humans can inform the design of automatic algorithms as well as future studies evaluating their performance.

Categories and Subject Descriptors

H.5.m. [**Information Interfaces and Presentation (e.g. HCI)**]: Miscellaneous

General Terms

Design

Keywords

tag clouds, structured, semantic, user study, visualization

AVI' 14, May 27 - 29 2014, Como, Italy
 ACM 978-1-4503-2775-6/14/05... $15.00.
http://dx.doi.org/10.1145/2598153.2598155

1. INTRODUCTION

Tag clouds are popular features on social media web pages. They are used to provide an overview over the content of a page or to support the user in searching or browsing activities. Tag clouds are a visual depiction of a set of words. The words are arranged spatially according to some layout criteria. The displayed terms are chosen by some rationale such as term frequency or other approximations of term importance. They can be extracted from a document or be selected from a set of user-generated tags that documents are annotated with in many social media applications.

In common *unstructured tag clouds*, the terms are arranged randomly or in alphabetic order, either in line-wise or free placement. In recent years alternative arrangements have been proposed that organize the tags in a cloud according to some similarity measure such as cooccurrences. It is supposed that such *structured tag clouds* come with improvements over the unstructured ones (cf. [6, 8, 17, 11]). However, only few efforts have been made to evaluate if this assumption is true. At the same time, many alternatives for automatic algorithms for generating structured tag clouds have been proposed, which are often based on more or less arbitrary choices on the selected term relations and the employed layout criteria. In this paper, we take an orthogonal, so far unexploited approach to first research how humans perform the task of generating a structured tag cloud. With this approach we follow the method suggested by van Ham and Rogowitz [20] that researched important criteria for layouting nodes in a network diagram by asking users to arrange the nodes in a way the relations in the data are captured best. Note that we do not assess the performance of structured tag clouds compared to unstructured ones nor do we propose a new algorithm. But we are able to reveal certain criteria that are shared among the layouts of our participants and that can inform the design of automatic algorithms, resulting in tag cloud arrangements similar to human ways to structure tags.

In summary, the goal of this study is to examine how humans structure user-generated tags when being told that the resulting tag cloud should provide a quick overview of the retrieval result for a tag search in the social bookmarking system BibSonomy (`http://bibsonomy.org/`). We deliberately focus on one specific task (gaining an overview, also called impression formation or gisting [17]), because it has been shown that the task that the cloud is used for (e.g., searching, browsing, gisting) may also effect what design is appropriate [19, 14]. Being aware that no layout will be equivalent to the next (because not only one possible solution exists that could be treated as a gold standard), we put special emphasis in the study on inquiring into the criteria that the participants aimed at

when layouting the cloud.

2. RELATED WORK ON STRUCTURED TAG CLOUDS

Koh et al. [10] introduced ManiWordle that allows a user to customly manipulate a tag cloud layout to reflect the desired structure. Approaches that propose automatic algorithms for generating structured tag clouds include [7, 1, 16, 9, 6, 13, 5]. Most of them approximate what they call the underlying "semantics" or "meaningful relations" between the tags with a cooccurrence-based measure. Exceptions include [12] that builds the cloud on subsumption relationships and additionally exploits Wikipedia concepts or [19] one of whose approaches uses WordNet to compute the relatedness of the tags. Almost none of the papers evaluates the generated structures in a user study, thus leaving it open if the proposed approach is not only technically sound but an improvement over unstructured tag clouds that can be intuitively read.

Regarding the layout, i.e., the visual representation of the determined term relations, many different variants have been proposed, e.g., line-wise representations with one line per cluster [7] or with 2D clusters whose terms are spread across multiple lines [18]; network representations [9, 6]; tag hierarchies [5]; 2D arrangements in which closeness reflects the relatedness of terms [21, 4, 15], some of which emphasize clustering structures [3, 1]; and SOM-based visualizations [16]. Only few of these suggested visual representations were evaluated with respect to their suitability and appropriateness for the specific task.

While already much research has been conducted on evaluating the performance of *unstructured* tag clouds, only few publications report on user studies that focus on (semantically) structured clouds. One of them is [19] that compares line-wise arrangements of alphabetical and random order to arrangements based on linguistics (exploiting WordNet) and folksonomy-based ones (using the getrelated-function of the Flickr API to determine term relations). In the tag cloud, semantically related tags are put close to each other but can be scattered over multiple lines. The authors could show that in a search task where a specific tag had to be identified the folksonomy-based layout was second best after the alphabetical one and was preferred by about half of the users for a general search task (identifying a tag which belongs to a specific topic). In a follow-up eye-tracking study [18] two different search strategies were identified (chaotic and serial search patterns) but no trend with respect to what strategy is used with what layout could be determined. The authors state that they "expect further advancement on semantic arrangements with more elaborate procedures" [19] and that the results might be different for other tasks such as browsing.

Lohmann et al. [14] compare alphabetically sorted tag clouds to circular and clustered ones. The arrangement of the alphabetically sorted tag cloud is line-wise. In the circular and clustered version, the tags are arbitrarily placed but the cloud is kept as compact as possible which is why the clusters are nearly not spatially separated from each other. One of the results of the study was that the clustered layout is overall preferred for finding tags that belong to a given topic but also that not all participants recognized the special arrangement. Furthermore, they recorded that for different user tasks different arrangements scored best.

Gou et al. [6] built a hierarchical clustering structure that is represented with a tag network. Each cluster can be drilled down on demand to reveal more details (subclusters). Besides evaluating if the tags that their algorithm considers related are also considered related by humans, they also conducted an experiment to find out if their built hierarchies resemble the ones that humans would build which was found to be true in most cases.

Further papers whose main focus is not on the evaluation of structured tag clouds but rather on the evaluation of a proposed visualization technique include [3] that compares its introduced representation of overlapping clusters in a task-oriented approach with common tag clouds. Tasks range from search tasks over identifying relations or judging tag relatedness. Knautz et al. [9] evaluate their whole system rather than the tag clouds as such but also come to the conclusion that "tag clusters are perceived as more useful than tag clouds, are much more trustworthy and significantly more enjoyable". Finally, [1] presents an approach that builds on hierarchical clustering in which the topics are separated spatially and by color. The system is compared to the Delicious interface in several undetermined browsing scenarios regarding three usability criteria. The study results "indicate enhanced support and user experience for the new interface".

To the best of our knowledge, so far no study exists that researches how humans address the task of generating a structured tag cloud.

3. METHOD

The purpose of this study was to find out what characterizes the layout[1] of a structured tag cloud generated by a human. Therefore, the participants were asked to arrange a set of user-generated tags from a social bookmarking system in a way that a quick overview over the content of a retrieval result can be gained. In the subsequent interviews special attention was given to the criteria underlying the applied layout strategy. The layout criteria of interest were specified in advance and were systematically inquired during the interview. A pilot study with four participants was conducted in the preparation phase which allowed us to iteratively improve the design until we were sure that the approach would yield expressive and reliable results.

3.1 Participants

We conducted the study with twelve participants that are all members of the German Institute for Educational Research and Educational Information (DIPF) in Frankfurt/Main, Germany and volunteered to participate in the study. Six of them are female and six male, ages ranged between 20 and 60. None of them participated in the pilot study. In terms of professional background, the participants can be clustered into three main groups: five of them are information specialists (documentalists), three have their background in information management, and the remaining four are trained in computer science or related areas (computational linguistics, mathematics). Thus, all participants can be considered experts in making information accessible but address the problem from different perspectives.[2] All of them have at least *used* Web 2.0 services such as Delicious, BibSonomy, Flickr, Wikipedia, Last.fm, etc., 83% have also contributed themselves to at least one of them.

3.2 Resources

To conduct the task, the participants were provided with 80 paper cards, each with a tag written on it. Those tag cards had to be ar-

[1]We use the term *layout* to refer to a (semantic) organization of a tag set. We are interested in the criteria relevant for structuring the terms (e.g., term relations taken into account) and not in how the terms can be most aesthetically or compactly arranged.

[2]Note that we do not distinguish between the three groups in the study because with only 3-5 participants per group interpersonal differences may be larger than the ones between the different groups. Consequently, we focus on the criteria that is *shared* by all participants rather than exploring the distinctions.

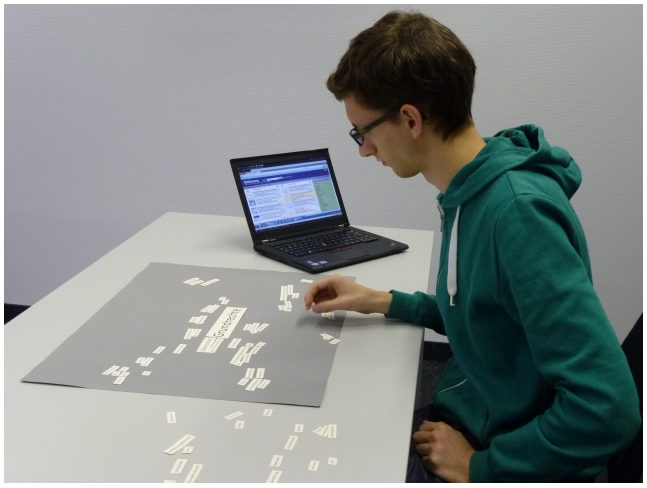


Figure 1: Setup of the user study.

ranged on a rectangular sized cardboard. As an initial setup the tag cards were placed in random order next to the cardboard. In addition, the participants were allowed to use the BibSonomy interface that the data was taken from and our Interactive Filtering Tool as information resources. The latter was specifically designed for the study to facilitate looking up cooccurrence relations between the tags. The tool displays a file in which all tag sets of the retrieval result are listed (one line per bookmarked webpage). The list can be filtered for lines that contain a certain tag to get an idea of the context the tag was used in.

3.3 Datasets

To avoid any bias that may be caused by special characteristics of a specific dataset, we retrieved data for three different query terms. A dataset is composed of the 80 most frequent tags of a BibSonomy retrieval result when querying the database for all web pages that were tagged with a specific query term. The query terms used in the study are 'Afrika' *Africa*, 'Kunst' *art*, and 'Grundrechte' *civil rights*. Our goal was to utilize data sets that can be made sense of with general knowledge which was confirmed by the participants after conducting the task. The font size of a term was mapped to its frequency in the dataset. For the datasets 'Afrika' and 'Grundrechte' the lowest frequency of selected terms was 2, for the 'Kunst' dataset it was 8. An offset was added to all frequencies to ensure that the smallest font size is 14. Furthermore, we had to downscale the font size of the very frequent query term 'Kunst', because it otherwise would have been too big to handle. Table 1 contains further statistics of the data. Each participant was given only one of the datasets, so in total each data set was worked on by four participants.

Though we queried the database with German terms, two of the datasets contained some English terms. To avoid any potential problems with language or unknown terminology, the participants were told that they could ask for definitions of the terms (no matter what language) if necessary.

3.4 Procedure

In the following, we detail the procedure of the study. On average the introduction took 15-20 min, completing the task about 30 min, and the following post-task interview about 30 min.

	Grundrechte	**Afrika**	**Kunst**
Mean value frequency	16.81	16.94	33.44
Standard deviation	8.37	7.92	37.94
Minimum font size	14	14	14
Maximum font size	86	84	255
# Web pages retrieved	74	72	837

Table 1: Dataset statistics

Pre-study questionnaire.

The pre-study questionnaire collected information about the participant such as experience with Web 2.0 services and professional background.

Introduction to the study.

First, an introduction to the social bookmarking system BibSonomy was given to ensure that all participants had enough background knowledge to understand the context of the task. Afterwards, the task and goal of the study were motivated by discussing different means to get an overview over a BibSonomy retrieval result. This was followed by an introduction to the task itself including an explanation of how the datasets were generated from BibSonomy.

The task involves arranging tags in a meaningful way. We were interested to find out if humans prefer to use lexical-semantic relations (such as synonymy, hypernymy / hyponymy, meronymy / holonymy), cooccurrence-based relations, semantic associations or a mixture of all of them. To ensure that the participants are aware of all those relations and also know what they mean, we explained all relations to them using a special demo dataset. Furthermore, the participants were informed that they do not have to employ all terms. We further pointed out that the layout can be compact or loose whatever seems best for fulfilling the task.

The participants were also introduced to both tools (see Sec. 3.2) before the start of the task. During the task, a notebook with the tools opened was placed next to the cardboard being available at all times. However, it was set at liberty to the participants if they use the tools during the layouting process or not. Figure 1 gives an impression of the setup.

Finally, the participants were informed about what data is collected and how this is done and asked for consent. The introduction procedure followed a detailed plan to ensure that all participants get the same information.

Completion of the task by the participant.

During the task completion, the researcher conducting the study observed the process but did not interrupt or intervene. The participants were encouraged to comment on their thoughts, decisions and rationales already during the layout process (think-aloud protocol). The task completion time was not restricted and on average 30 mins were used for the layouting process (min: 20 mins, max: 40 mins).

Post-task interview.

The interview started with a detailed explanation of the layout by the participant. Where necessary, the interviewer asked clarifying questions until she was sure that she had understood the layout and the intentions behind it.

This was followed by a detailed structured interview. All criteria that we were interested in were systematically inquired. Some questions targeted at quantitative values or answers to multiple choice questions, whereas others collected free answers that pro-

vided information about the explanations, intentions and rationales behind layout decisions. The following issues were addressed:

- *Resources*: Which of the resources (BibSonomy, Interactive Filtering Tool) were used and how often? (recorded by the interviewer & logged) What were they used for?
- *Sorted out terms*: Which terms were sorted out? (photographed) What was the rationale behind this?
- *Start terms*: Which terms made the starting point (registered by the interviewer) and why those?
- *Layout strategy*: What was the general layout strategy? Was it clear from the beginning or did it develop while conducting the task?
- *Term relations*: Which relations were taken into account and how did they affect the layout?
- *Cooccurrences*: Did they inform the design? What are the participants' opinions on employing cooccurrences?
- *Difficult terms*: Were there difficult to integrate terms and how was dealt with them?
- *Higher-level structures*: What higher level structures are present in the layout (clusters, hierarchies, misc)? How many manifestations of this kind are included? (The participants were asked to mark them in a print-out of a picture of the final layout.)
- *Global structures*: Do global structures exist? Is the distance between the higher-level structures meaningful?
- *Lessons learned*: If the participant could start over again, would he/she change anything?
- Did the participant feel comfortable with the dataset?

3.5 Data Collected

The interviewer took notes during the task conduction regarding tool usage, terms started with, comments given by the participant during the layout process and other observations that seemed interesting or required further investigation during the subsequent interview. Query terms entered into the Interactive Filtering Tool were logged. Detailed notes were also taken during the post-task interview. The interview was additionally recorded (except for 2 cases in which the participants objected against being recorded). Finally, pictures of the layout were taken at intermediate steps of the layout process as well as at the end to record the final result of the layouting process.

3.6 Analysis of the data

Partly, our collected data could be analyzed quantitatively such as counting the number of terms sorted out, the number of clusters built, noting whether a term relation was taken into account or not etc. Where this was not the case but the free-form statements of the participants had to be interpreted, emergent coding was applied. Thereby, the coding categories were determined independently by the PostDoc researcher who also conducted the interviews and a student assistant. Afterwards, the derived categories were discussed and refined until an agreement was reached. Similarly, the coding itself was also conducted independently by the two researchers and afterwards compared and where necessary discussed. In case of disagreements, usually a consensus could be reached quickly in the discussion except for one case in which the whole procedure was repeated with additional data from the recordings.

TN-ID	Synonymy	German/ English	Hyponymy/ Hypernymy	Meronymy/ Holonymy	Semantic association
1	N	O	H	H	N
2	N	-	N	x	N
3	N	N	N	N	N
4	O	O	O	O	N
5	N	-	N/S	N	N
6	S	S	N	x	N
7	O	O	N	N	N
8	O	-	O	O	N/O
9	N	N	N/H	N	N
10	N	N	N	N	N
11	O	O	N/O	x	N
12	N	-	N	x	N

Table 2: Linguistic relations and how they were taken into account. N = next to each other, O = sorted out, S = separated, H = below/above (hierarchy), x = not taken into account, - not present in dataset

4. RESULTS

In the following we summarize the results of the user study based on the resulting layouts and the answers to the questions of the post-task interview.

4.1 Linguistic Relations

In the post-task interview, we asked the participants for each of the term relations[3] that were introduced before the task if they took it into account and how it influenced their layout. Table 2 provides details about the usage of the linguistic relations. The German/English column refers to German-English term pairs where the one is a translation of the other (e.g., *Wissenschaft - science*).

We can observe that either one of the synonyms (terms with same or similar meaning) was sorted out or both terms were put close to each other. As an exception to the rule, one participant left both terms in, but put them at different places if they fitted in several clusters. It is striking that almost all participants treated English/German term pairs the same as all other synonyms.

Surprisingly, the hierarchical relationship that is inherent to hypernymy / hyponymy (superordinate / subordinate term) was rarely reflected in the participants' layouts. Mostly, the terms were treated similar to synonyms and simply put next to each other or sorted out to avoid redundancy. Meronymy / holonymy (part-of) relations were not taken into account at all or treated the same as hypernymy / hyponymy. Some participants explicitly stated in the post-task interview that they did not distinguish between those two relations at all.

We also asked our participants what role semantic associations played in the layout process. Such term pairs belong to a common concept or topic but are not necessarily related to each other in terms of the lexical-semantic relations outlined above. The terms *refugee, asylum-seeker, immigration office, refugee camp* could be considered an example for a cluster of semantically associated terms. It is easy to see that all those terms semantically belong to the same topic but except for the term pair *refugee* and *refugee camp*, which also could be classified as a meronymy relation, do not comply with any of the lexical-semantic relations discussed above.

[3]We use *linguistic relations* as a superordinate term for *lexical-semantic relations* (such as synonyms, hypernyms/hyponyms, meronyms/holonyms) and *semantic association*. *Term relation* is used if all kinds of relations between terms are referred to including, e.g., *cooccurrence-based relatedness*. *Semantic relatedness* is used synonymously to stress that the relation is constituted by some (unspecified) semantics.

Semantic associations were taken into account in all layouts. When asked what term relation most influenced their layout, eight participants stated that semantic associations were the most important factor. Three other participants explained that not a single but rather a combination of relations, namely semantic associations together with hypernymy and meronymy is what their layout was built upon (interestingly, all three participants belong to the group of information specialists). Finally, one participant organized the terms along a timeline (using the timestamps in BibSonomy) and consequently reported "time" as the most important factor for the layout.

4.2 Role of cooccurrences and resource usage

In the course of investigating the role of the different term relations, we also asked the participants if cooccurrence-based relations were taken into account in their layout which was the case for one third of our participants. Independent of their answer, all participants were also asked what they think in general about the usage of cooccurrences for the task, in what respects they consider this relation as beneficial, but also what doubts or concerns they have against basing layouts on cooccurrences. The data collected was enriched with statements that the participants made about their usage of the BibSonomy interface and about the Interactive Filtering Tool (IFT) that could be used for investigating cooccurrence relations. In total 50% of the participants made use of the IFT and 25% of the BibSonomy interface. In the following, we report on the valuations of the participants.

Most often (five times) the participants mentioned that they consider cooccurrences useful to investigate the meaning of unknown terms by looking up their usage context. Four participants regarded cooccurrence relations as a good means to take the collection bias[4] into account (i.e. to find out in which context a term is used in this specific collection if it has multiple senses or could be used in different thematic contexts). Further mentions referred to the value of cooccurrences for generating ideas to get started (mentioned twice) and for finetuning the result (mentioned once).

On the other hand, four participants worried that a layout primarily based on cooccurrences might be difficult to understand. The same number of participants expressed at least doubts that cooccurrence relations reflect their notion of semantics. Finally, two participants stated that they did not know how to incorporate cooccurrence relations into their layout or that including this relation would not fit to their developed layout strategy.

Besides, we had one participant who indeed aimed at basing her layout on cooccurrences in order to reflect the collection bias as well as possible. She noted that key for making her approach work was to sort out all terms that are general in the sense that they (theoretically, not based on the collection!) could be added everywhere, because otherwise her layout would become difficult to read. Consequently, only 41 of the 80 terms were left in and, for instance, all terms denoting types of media (a typical meta data cluster, see Sec. 5) were sorted out.

4.3 Higher-level and global structures

We asked each participant after the study to mark higher-level structures in a print-out of their final layout. The predominant higher-level structures were clusters. Five participants implemented also subclusters in one or more clusters. On average 5.8 clusters existed per cluster-based layout (if the lowest subclustering level is taken into account 7.6 clusters). Five participants allowed individual terms that do not belong to any cluster or interpreted them simply as single-term clusters.

Two-thirds of the clusters consisted of ten or less terms (about half of which had 5 or less terms), whereas 8% had more than 20 terms. Interestingly, most of the larger clusters had an inner substructure; i.e., they were themselves organized in subclusters, a hierarchy, or according to semantic relatedness.

Next, we investigated the role of global structures. A global structure exists if the arrangement of the higher-level structures (clusters) is meaningful (i.e., changing the positions of the clusters would destroy an underlying semantics). The interview revealed that all participants had in some way or another a global structure. However, whereas this structure was central in some layouts, in other cases it affected only part of the arrangement. In most cases, the global structure was constituted by arranging semantically related clusters next to each other. Besides, two participants implemented a global hierarchy of clusters and one arranged them along a timeline (showing event clusters only). Another participant arranged the clusters along thematic rays that start in the query term. Here the distance to the query term encoded the degree of relatedness of the term cluster to the query term. Finally, half of the participants stated that the distances between the clusters are meaningful and can be compared to each other.

4.4 Terms sorted out

With respect to the number of terms sorted out, our participants split up into two contrasting groups: eight participants sorted out only few terms (8 or less) and the remaining four participants each left out 36 or more terms. Thereby, two different strategies became apparent: The ones that sorted out many terms deliberately selected those terms that they thought would be important for giving an overview and left the remaining ones out. They consequently tried to remove all redundancies (especially synonyms or singular/plural but also hypernyms or meronyms, see also Table 2). Furthermore, three of the four stated that they had also sorted out terms that seemed too specific or general for providing information in an overview. In contrast to this, the other participants aimed at integrating all terms and only left out the ones they could not add because the term was unknown, was difficult to integrate given their layout strategy and terms for which they did not know how to associate them with the topic.

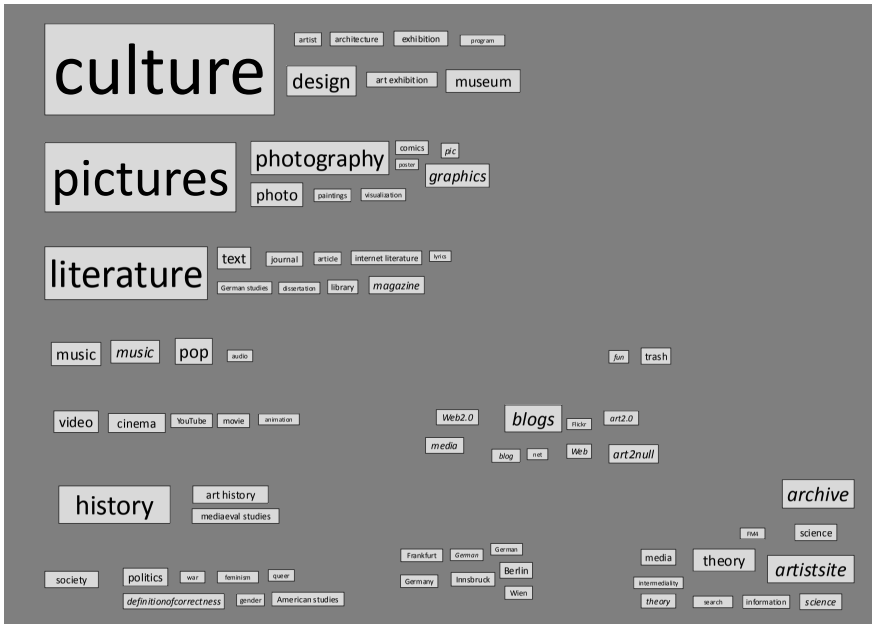
4.5 Resulting layout

Figure 2 shows four examples for human-generated layouts. To ensure interpretability also for non-German speaking readers, we translated all terms into English and display reproduced images, keeping the positions of the different terms as close as possible to the photographed cloud.

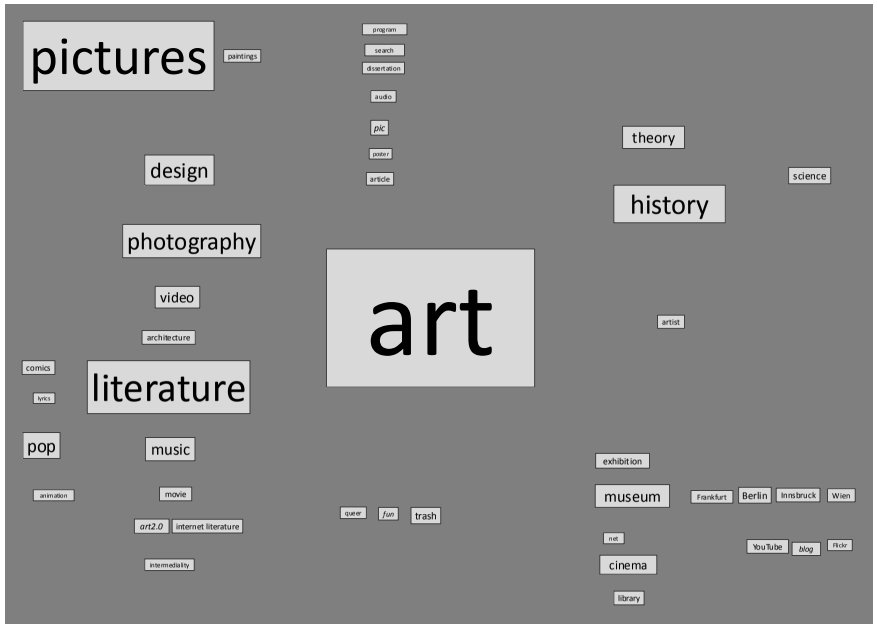
We asked all participants whether their layout strategy was clear from the beginning or developed while conducting the task. Most participants responded that the latter was true. However, two documentalists and two information management professionals pointed out that they had to deal with similar or related tasks in the past in the course of their studies or in their everyday working life which affected their strategy.

As expected all generated layouts were unique which is why we designed the study in a way that the post-task interview revealed the commonalities of the underlying layout criteria. Nevertheless, we additionally inspected the resulting structured tag clouds for shared concepts.

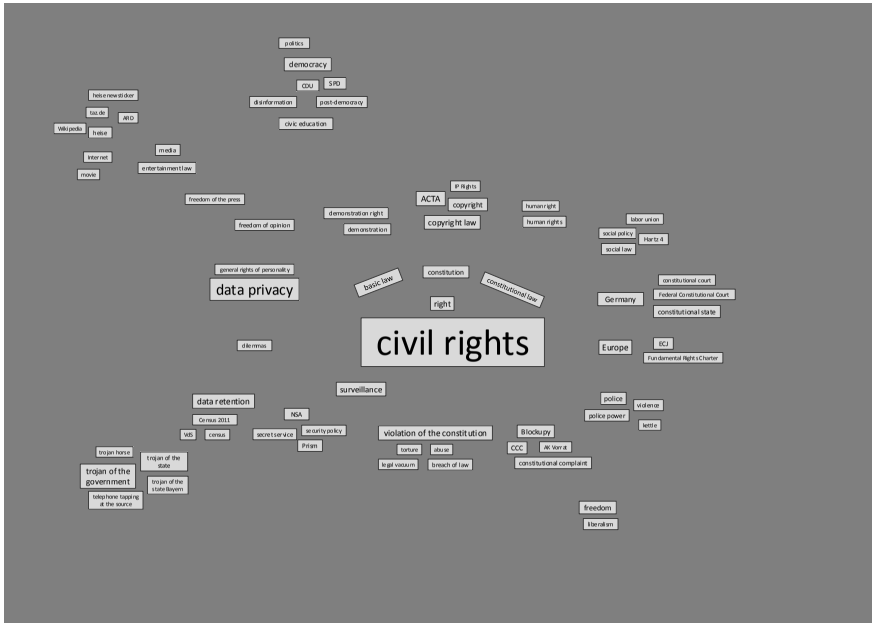
[4]The term "collection bias" refers to the fact that the topic coverage of the document collection under investigation might not fully comply with common associations with the query term. For instance did our Africa dataset not contain documents about vacation or wild life. Consequently, the strongest associations with the terms might not be the most appropriate ones for the specific collection.



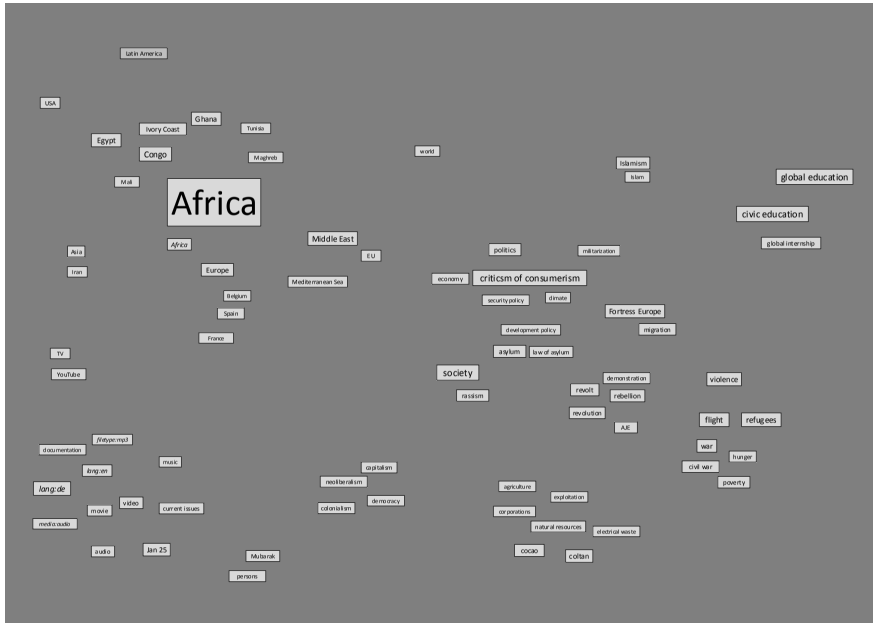
(a) Kunst (art) dataset (1)



(b) Kunst (art) dataset (2)



(c) Grundrechte (civil rights) dataset



(d) Afrika (Africa) dataset

Figure 2: Samples of human-generated layouts reproduced and translated into English. Terms in italics were already English in the original dataset used for the study.

The 'Kunst' *art* dataset was clearly the dataset with most consensus. Not only did the participants all build definite cluster structures that they could assign labels to but they also had many shared cluster topics such as places of art, theory and science terms, types of media, or art genres. The layouts for the 'Grundrechte' *civil rights* dataset were characterized by their large variety in terms of global structures. The variants ranged from association chains over rays with clusters and a timeline to a complex clustering structure including subclusters. The participants that structured the 'Afrika' *africa* tags all worked with clustering structures. However, in contrast to the 'Kunst' dataset a wide variety of topics for the clusters can be observed. Possibly, this was due to the complex thematic structure of the dataset, as some participants explicitly pointed out that they were torn between different alternative structurings and that they see interconnections between the cluster topics.

Part of the observed larger variety for the two latter datasets might be explained by the fact that they were built from a smaller data base than the 'Kunst' dataset and consequently also less frequent and more specific terms and topics were included. Even more remarkable it is that even those datasets had topics that were recognized and considered in the layout by most participants such as clusters of geographical terms and media types that could be found in many tag clouds. Those associations were described as "obvious" or "natural" in the interview.

4.6 Miscellaneous

Regarding the terms chosen to start with, the most frequently mentioned reasonings were to use tags with a large font size first as well as the ones which seemed strongly associated to the query term.

Difficult to integrate terms were often placed next to terms that they were (subjectively) most associated with or where they neatly fitted into the layout. Four participants reported that they tried to find out in which context they were used most in the given collection. Overall, the participants did not seem to be too much worried about single terms being placed at suboptimal positions.

5. DISCUSSION & CONCLUSIONS

In the following we reflect on the results of the study, identifying lessons learned that will inform design decisions of algorithmic approaches or future studies.

Semantic Associations

Approaches in related work often postulate that the layout of a tag cloud should be built on "semantics" or more generally "a meaningful relation" but do not further detail on how this is defined. Our study reveals that **semantic associations are the main criterion for human layouters to build their overall structure on**. Semantic associations are not a clearly definable concept and therefore no absolute ground-truth exists. Nevertheless, in our study there were certain groupings that showed up over and over again such as terms related to theory and science, geographical terms or types of media. These clusters describe concepts that are not specific for the query term or for a certain subarea of the search topic but could be used in different contexts and therefore could be considered as meta data. (One of our participants denoted them as "secondary clusters".)

Lexical-semantic relations

All participants were able to state how the lexical-semantic relations such as synonyms, hypernyms, meronyms etc. were integrated in the layout. However, many had to first inspect their layout in detail during the interview to be able to answer the question which suggests that those relations were not consciously taken into account. This assumption is supported by the fact that most participants did not organize the hypernym / hyponym relations hierarchically but put them close to each other recognizing them as "related". Eventually all examined lexical-semantic relations can also be considered semantic associations (while not all semantic associations are also lexical-semantic relations). Knowing this, the lexical-semantic relations **can be a further means to identify semantic assocations in an automatic layout method**, e.g., by leveraging lexical resources such as WordNet (`http://wordnet.princeton.edu/`). Furthermore, lexical-semantic relations turned out to be **the basis for determining redundant terms** that can be sorted out to reduce the number of terms in the cloud if desired.

Cooccurrence relations

In our study, two tags are considered as cooccurring if they are assigned to the same bookmarked webpage. The most frequent purpose that the participants considered such cooccurrences useful for was to get an idea of the meaning of a term by looking at the context it was used in. The method could be adopted by an automatic algorithm **to integrate uncommon terms** such as neologisms or informal but frequent community slang into the layout.

Besides this, cooccurrence relations suggest themselves as a good means to take the collection bias into account. Yet, only four of our participants recognized this as a benefit of cooccurrences, whereas seven participants expressed doubts if cooccurrence relations can capture their own notion of semantic relatedness or worried that building the layout on cooccurrence relations might result in a difficult to interpret tag cloud. This is especially worth considering because most approaches proposed so far in literature deeply rely on cooccurrence relations. **In general, the participants' structures seemed rather influenced by common knowledge of what typically relates to each other than by the collection bias.** This is in line with the comments of a couple of participants that described their strategy of grouping thematic terms as putting next to each other what was related in the recent news coverage.

Are the doubts of our participants regarding the suitability of cooccurrence-based relations justified? The fact that not all participants were familiar with cooccurrence relations before the study and that cooccurrences might not be a natural means for a human to build a layout on impedes reliable conclusions. Furthermore, the study did not investigate the performance of the generated clouds and therefore cannot answer the question if disregarding the collection bias misleads or eases the interpretation and how the usage of cooccurrences effects the ease of interpretation and usability. But all in all, the preliminary findings suggest that it could be **worth to evaluate the effect of the common practice to build layouts solely on cooccurrence-based relations in more detail**.

Higher-level structures

With respect to higher-level structures we could observe that **small clusters are preferred over large ones** (see details in Sec. 4). This is in line with Begelman et al. [2] that recommend to choose a clustering algorithm that produces rather small clusters. In addition, we can learn from the study that **where larger clusters cannot be avoided, they should be further structured internally** to retain clarity.

Global structure

All of our participants also implemented a global structure (e.g., semantically meaningful arrangements of the clusters) which leads us to believe that this is a **concept that can be understood** by users and could therefore be made use of in an algorithmic approach as well.

Font size
We did not analyze the effect of font size, however, studies on unstructured tag clouds revealed that the font size effects the perception of a tag cloud, e.g., that larger terms attract more attention. It stands to reason that this effect also influences the perception of a *structured* tag cloud. Hearst and Rosner [8] postulate that larger terms should bear the main message. This is supported by the statements of three participants of the study that expressed the wish to change the font size in a way that the most central term of their cluster is also the largest one. On the other hand they also reasoned about the need to see the term frequency which hints at the importance of the term within the specific collection. Possibly, an additional visual variables such as color could be used to encode both aspects.

In future work we intend to develop an automatic layout algorithm that implements the discovered criteria and to further investigate the advantages of structured tag clouds.

6. ACKNOWLEDGMENTS

Many thanks to Christoph Schindler for his support during the design phase of the user study and to Catherin Mohr who helped with the emergent coding. This work was partially funded by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806.

7. REFERENCES

[1] H. Aras, S. Siegel, and R. Malaka. Semantic Cloud: An Enhanced Browsing Interface for Exploring Resources in Folksonomy Systems. In *Proc. of the Workshop on Visual Interfaces to the Social and Semantic Web*, VISSW2010, 2010.

[2] G. Begelman, P. Keller, and F. Smadja. Automated Tag Clustering: Improving search and exploration in the tag space. In *Proc. of the Collaborative Web Tagging Workshop*, 2006.

[3] Y.-X. Chen, R. Santamaría, A. Butz, and R. Therón. TagClusters: Semantic Aggregation of Collaborative Tags beyond TagClouds. In *Proc. of the 10th Intern. Symp. on Smart Graphics*, SG '09, pages 56–67, 2009.

[4] W. Cui, Y. Wu, S. Liu, F. Wei, M. Zhou, and H. Qu. Context preserving, dynamic word cloud visualization. *IEEE Computer Graphics and Applications*, 30(6):42–53, 2010.

[5] P. Gambette and J. Vèronis. Visualising a Text with a Tree Cloud. In H. Locarek-Junge and C. Weihs, editors, *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 561–569. Springer, 2010.

[6] L. Gou, S. Zhang, J. Wang, and X. Zhang. Tagnet: Supporting the Exploration of Knowledge Structures of Social Tags with Multiscale Network Visualization. *Intern. J. of Advanced Intelligence*, 3(1):67–93, 2011.

[7] Y. Hassan-Montero and V. Herrero-Solana. Improving Tag-Clouds as Visual Information Retrieval Interfaces. In *Proc. InSciT 2006*, 2006.

[8] M. A. Hearst and D. Rosner. Tag Clouds: Data Analysis Tool or Social Signaller? In *Proc. of the 41st Annual Hawaii Intern. Conf. on System Sciences*, HICSS '08. IEEE Computer Society, 2008.

[9] K. Knautz, S. Soubusta, and W. G. Stock. Tag Clusters as Information Retrieval Interfaces. In *Proc. of the 2010 43rd Hawaii Intern. Conf. on System Sciences*, HICSS '10, pages 1–10. IEEE Computer Society, 2010.

[10] K. Koh, B. Lee, B. Kim, and J. Seo. Maniwordle: Providing flexible control over wordle. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1190–1197, Nov. 2010.

[11] B. Y.-L. Kuo, T. Hentrich, B. M. Good, and M. D. Wilkinson. Tag clouds for summarizing web search results. In *Proc. of the 16th Intern. Conf on World Wide Web*, WWW '07, pages 1203–1204. ACM, 2007.

[12] K. Lee, H. Kim, C. Jang, and H.-J. Kim. Folksoviz: a subsumption-based folksonomy visualization using wikipedia texts. In *Proc. of the 17th Intern. Conf. on World Wide Web*, WWW '08, pages 1093–1094. ACM, 2008.

[13] R. Li, S. Bao, Y. Yu, B. Fei, and Z. Su. Towards effective browsing of large scale social annotations. In *Proc. of the 16th Intern. Conf. on World Wide Web*, WWW '07, pages 943–952. ACM, 2007.

[14] S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration. In *Proc. of the 12th IFIP TC 13 Intern. Conf. on Human-Computer Interaction: Part I*, INTERACT '09, pages 392–404. Springer-Verlag, 2009.

[15] F. V. Paulovich, F. M. B. Toledo, G. P. Telles, R. Minghim, and L. G. Nonato. Semantic wordification of document collections. *Comp. Graph. Forum*, 31(3pt3):1145–1153, 2012.

[16] A. Pérez García-Plaza, A. Zubiaga, V. Fresno, and R. Martínez. Reorganizing clouds: A study on tag clustering and evaluation. *Expert Syst. Appl.*, 39(10):9483–9493, 2012.

[17] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen. Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '07, pages 995–998. ACM, 2007.

[18] J. Schrammel, S. Deutsch, and M. Tscheligi. Visual Search Strategies of Tag Clouds - Results from an Eyetracking Study. In *Proc. of the 12th IFIP TC 13 Intern. Conf. on Human-Computer Interaction: Part II*, INTERACT '09, pages 819–831. Springer-Verlag, 2009.

[19] J. Schrammel, M. Leitner, and M. Tscheligi. Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '09, pages 2037–2040. ACM, 2009.

[20] F. van Ham and B. Rogowitz. Perceptual organization in user-generated graph layouts. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1333–1339, Nov. 2008.

[21] Y. Wu, T. Provan, F. Wei, S. Liu, and K.-L. Ma. Semantic-preserving word clouds by seam carving. In *Proc. of the 13th Eurographics / IEEE - VGTC Conf. on Visualization*, EuroVis'11, pages 741–750, 2011.